Label Errors in BANKING77

Cecilia Ying
Queen's University
Smith School of Business
y.ying@queensu.ca

Stephen Thomas
Queen's University
Smith School of Business
stephen.thomas@queensu.ca

Abstract

We investigate potential label errors present in the popular BANKING77 dataset and the associated negative impacts on intent classification methods. Motivated by our own negative results when constructing an intent classifier, we applied two automated approaches to identify potential label errors in the dataset. We found that over 1,400 (14%) of the 10,003 training utterances may have been incorrectly labelled. In a simple experiment, we found that by removing the utterances with potential errors, our intent classifier saw an increase of 4.5% and 8% for the F1-Score and Adjusted Rand Index, respectively, in supervised and unsupervised classification. This paper serves as a warning of the potential of noisy labels in popular NLP datasets. Further study is needed to fully identify the breadth and depth of label errors in BANKING77 and other datasets.

1 Introduction

NLP researchers and practitioners use standard benchmark datasets in the selection, development, and comparison of advanced NLP methods. The use of standard benchmarks enables an apples-to-apples comparison of competing methods, as well as an evaluation of a method under different business scenarios.

Recently, researchers have proposed three promising intent classification benchmark datasets that are large (>10,000 instances) and include more than 50 unique intents: BANKING77 (cas), HWU64 (Liu et al., 2019), and CLINC150 (lar).

The aforementioned datasets have been used to evaluate pretrained transformers (Zhang et al., 2021b), density-based models (gon), few-shot learning (luo), open intent detection (Zhang et al., 2021a), and intent discovery (cha).

These benchmark datasets are hand-labelled by humans and their categorization can be subjective in nature. In addition, humans may make mistakes in the labelling process. As such, it is important to assess the accuracy of the human-given labels (Northcutt et al., 2021a).

Our recent experience with BANKING77 suggested that several labeling errors were present in the dataset. Using confident learning (Northcutt et al., 2021b) and our own cosine similarity methodology (Section 3.2), we found that over 1,400 (14%) of the 10,003 training samples may have been incorrectly labelled. Table 1 shows representative examples.

Using noisy labels to train and evaluate an intent classifier could have disastrous consequences. First, the classifier could incorrectly classify new utterances. Second, any performance measures would be based on mislabelled truth and therefore be inaccurate. Finally, researchers and practitioners may make an incorrect recommendation or conclusion for the downstream task-oriented conversational system.

In this paper, we investigate the potential label errors present in BANKING77. First, we provide background on BANKING77 in Section 2. In Section 3, we describe our methodology for determining potential label errors. We first use Confident Learning (Northcutt et al., 2021a) and identify over 900 potential label errors. Next, we design a methodology based on cosine similarity and identify an additional 500 potential label errors. In Section 4, we quantify the potential impacts of errors on a downstream NLP task. Finally, in Section 5 we conclude and outline future work.

2 Background

BANKING77 was created in 2020 by researchers at PolyAI[1] as part of their study on a new intent classifier using pretrained dual sentence encoders based on fixed Universal Sentence Encoders (Cer et al., 2018) and ConveRT (Henderson et al., 2020). The dataset is a single-domain intent detection

[1] github.com/PolyAI-LDN/task-specific-datasets/tree/master/banking_data

Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 139 - 143
May 26, 2022 ©2022 Association for Computational Linguistics

Similar utterances with different labels	
Utterance	**Label**
"How long will it take for me to get my card?"	card_arrival
"Can you tell me how long it takes for a new card to come?"	card_delivery_estimate
"Can you tell me the status of my new card?"	lost_or_stolen_card
"how many days processing new card?"	contactless_not_working
"Can you tell me when my money transfer will go through"	pending_transfer
"How long am I to wait before the transfer gets to my account?"	transfer_timing
"How long before a bank transfer shows up in the account?"	balance_not_updated_after_bank_transfer
Dissimilar utterances with the same label	
Utterance	**Label**
"How do I check security settings using the app?"	card_not_working
"I cannot seem to use my card."	card_not_working
"Can I use app to reset PIN attempts?"	card_not_working
"How do I check security settings on my card?"	card_not_working
"HOW LONG TO TAKE THE TIME TO SOLVE"	card_not_working

Table 1: Examples of potential label errors. The top portion shows utterances with similar intents assigned to different labels. The bottom portion shows examples of utterances with different intents assigned to the same label.

dataset, containing 10,003 annotated customer service queries over 77 intents related only to banking.

Many of the previously available datasets only included a small number of labels and contained a small number of utterances from many distinct domains. The authors believe that BANKING77—given its single-domain focus yet large number of intents—makes the intent detection task more realistic and challenging.

The authors also acknowledged that there are partially overlapping intent categories, and therefore, the intent detection system cannot rely only on the semantics of individual words to correctly categorize the utterance. However, they did not provide any specifics regarding the extent and impact of such overlaps.

3 Identifying Potential Label Errors

While implementing our own intent classifier on BANKING77, we noticed unexpectedly poor performance in several intent categories. We found that our classifier was confusing many of the labels. For instance, we found that up to sixteen “truth” labels were predicted as a single intent by our classifier. Similarly, one predicted intent included up to twelve truth labels. (Table 1 shows examples of such confusion.) While some prediction errors are expected, we were quite surprised at the level of confusion. We performed a preliminary manual investigation of labels and found that many utterances seemed to have the wrong truth label assigned. Also, we found that labels related to *"card"* or *"top_up"* have high similarities, as shown in Figure 1, making it difficult to select a distinct and unique label.

To further understand the extend of these potential label errors, we applied and compared two automated approaches: the Confident Learning framework, and a Cosine Similarity approach.

3.1 Confident Learning Framework

We replicated the Confident Learning (CL) framework (Northcutt et al., 2021b)[2], which produces a *label noise estimation* to find potential label errors, identified through the joint distribution of the noisy (given) labels and latent (unknown) labels to characterize class-conditional label noise.

We trained a LightGBM classifier on SBERT (rei) MPNet (Song et al., 2020) sentence embeddings. We used 10-fold cross validation to obtain out-of-sample predictions to identify potential label errors.

We found that 965 utterances, representing 75 of the 77 labels, may have potential label errors. Table 2 summarizes the top five labels with the highest number of possible errors. It is interesting to point out that utterances related to *"transfers"* or *"top_up"* labels appear to be most problematic.

3.2 Cosine Similarity Approach

The CL approach excelled at finding utterances that were identified as noisy within the same label. However, in our manual investigation, we also noticed that many utterances were semantically identical (e.g., “*Why hasn’t my transfer gone through*” and “*Why is my transfer still pending?*”) but were assigned different labels.

[2] https://github.com/cleanlab/cleanlab

Label	Potential Errors
transfer_not_received_by_recipient	32
balance_not_updated_after_bank_transfer	31
top_up_failed	24
top_up_reverted	24
pending_top_up	23

Table 2: The top five labels with potential errors from the CL framework.

Label	Potential Errors
card_arrival	42
getting_virtual_card	37
declined_card_payment	33
pending_top_up	33
top_up_reverted	30

Table 3: The top five labels with potential errors from our Cosine Similarity approach.

We created a method to find such utterances as follows. First, we calculated the pairwise cosine similarity (based on SBERT MPNET embeddings). Next, we identified pair of utterances that had similarity score higher than δ =0.85 but were assigned different labels.

We found that 590 utterances, representing 49 of the 77 labels, may have potential label errors. Table 3 summarizes the top five labels with the most conflicting labels assigned to similar utterances. Utterances related to *"card_arrival"* have the largest number of label disagreements.

We also noticed that two labels related to *"top_up"* have been identified by both approaches, indicating further investigation related to these two labels is needed. 127 of the 10,003 utterances were identified as potential label errors by both approaches, of which only 80 shared the same suggested correct labels.

4 Experiment Results

To illustrate the negative impact of the noisy labels on the performance of an intent classifier, we designed an experiment as follows.

First, we considered two versions of the BANKING77 dataset. The **original**, unmodified version, and a **trimmed** version whereby we removed all utterances with potential label errors identified by either the CL framework or cosine-similarity approach. Table 4 compares the statistics between the original and the trimmed version of the dataset.

	Original	Trimmed
Unique labels	77	77
Utterances	10,003	8,575
Terms	4,518	4,230
Tokens	119,530	103,776
Tokens per utterance	11.9	12.1
Mean term occurrence	26.5	24.5

Table 4: Statistics of the original and trimmed versions of the BANKING77 dataset.

Next, we built two intent classifiers, one supervised and one unsupervised, as follows. We obtained sentence embeddings for each dataset using SBERT and MPNet. We reduced the dimensionality of the embeddings using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) (n_components=20, n_neighbors=40).

In the supervised approach, we used LightGBM (n_estimators = 1000, learning_rate = 0.1, max_depth=4, num_leaves=15) to train two models. Using 5-fold cross validation, we measured each model's accuracy and F1-score.

For comparison, we used Agglomerative Clustering (n_clusters=77, affinity="euclidean", linkage="ward") as our unsupervised approach. We then measured five common clustering metrics: Adjusted Rand Index (ARI); Adjusted Mutual Information (AMI), Completeness, Fowlkes-Mallows, and Homogeneity.

Table 5 shows the results. We find that by removing utterances flagged as potential errors significantly improved the performance of the intent classifier according to all metrics. Notably, F1-score increased by **4.5%** in the supervised approach, and ARI increased by **8%** in the unsupervised approach.

5 Conclusion and Future Work

In this paper, we investigated potential label errors present in the popular BANKING77 benchmark dataset. We applied two automated techniques to identify potential label errors. First, we used the Confident Learning framework to find utterances based on class-conditional noise estimates. Sec-

Supervised Classifier			
LightGBM			
Metric	**Original**	**Trimmed**	**% Diff**
Accuracy	0.882	0.924	+4.5%
F1-Score	0.878	0.920	+4.5%
Unsupervised Classifier			
Agglomerative Clustering			
Metric	**Original**	**Trimmed**	**% Diff**
ARI	0.6344	0.6859	+8%
AMI	0.8333	0.8565	+3%
Completeness	0.8527	0.8735	+2%
Fowlkes-Mallows	0.6409	0.6909	+8%
Homogeneity	0.8392	0.8648	+3%

Table 5: Experiment results. We report various metrics on the original dataset, the trimmed dataset, and the difference between the two. ARI is the Adjusted Rand Index and AMI is the Adjusted Mutual Information.

ond, we developed our own cosine-similarity based technique to find utterances that are semantically similar but labeled differently. Together, these approaches identified over 1,400 utterances with potential label errors. A simple experiment showed that an intent classifier's performance can be improved by removing such utterances. F1-score increased by **4.5%** for the supervised classifier, and ARI increased by **8%** for the unsupervised classifier.

Given the importance of benchmark datasets in the development, evaluation, and selection of NLP techniques, it is important that the labels contain as few errors as possible. We would like to extend our work by developing an automated correction tool that can identify and fix label errors. We will also manually verify and correct errors in BANKING77, and it will serve as the ground truth for evaluating the performance of the automated correction tool. Furthermore, we will apply the methodology on other benchmark datasets such as CLINC150 and HWU64.

References

Density-Based Dynamic Curriculum Learning for Intent Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Virtual Event Queensland Australia.

Don't Miss the Labels: Label-semantic Augmented Meta-Learner for Few-Shot Text Classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online.

Efficient Intent Detection with Dual Sentence Encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, Online.

An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China.

Intent Mining from past conversations for Conversational Agent. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online).

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *arXiv:1803.11175*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and Accurate Conversational Representations from Transformers. *arXiv:1911.03688*.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking Natural Language Understanding Services for building Conversational Agents. *arXiv:1903.05566*.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426*.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021a. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv:2103.14749*.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2021b. Confident Learning: Estimating Uncertainty in Dataset Labels. *arXiv:1911.00068*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. *arXiv:2004.09297*.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021a. Deep Open Intent Classification with Adaptive Decision Boundary. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16).

Jian-Guo Zhang, Kazuma Hashimoto, Yao Wan, Ye Liu, Caiming Xiong, and Philip S. Yu. 2021b. Are Pretrained Transformers Robust in Intent Classification? A Missing Ingredient in Evaluation of Out-of-Scope Intent Detection. *arXiv:2106.04564*.

A Appendix - Similarities between labels

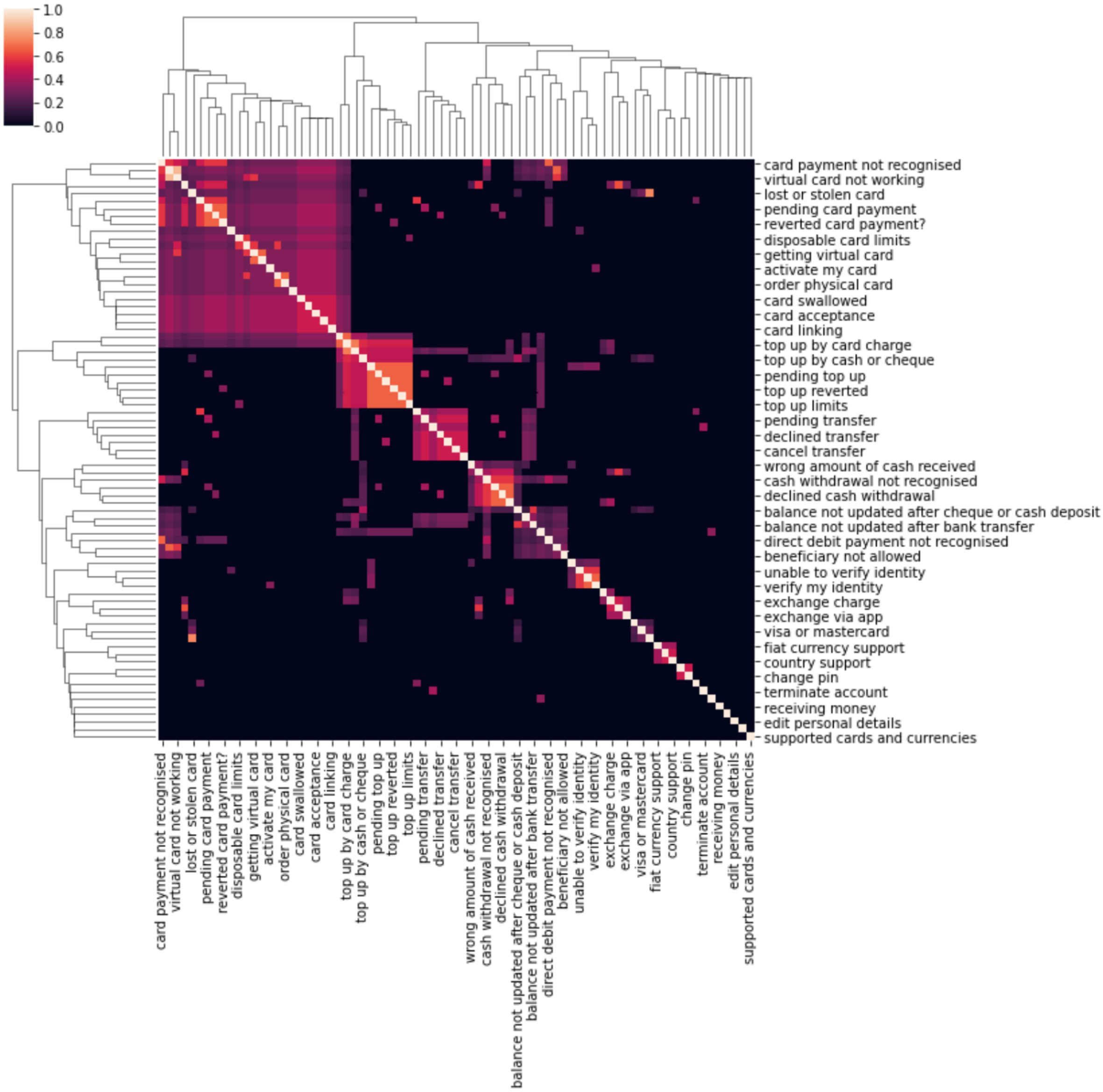

Figure 1: A heatmap of label similarities in the BANKING77 dataset, according to a simple word count. Labels are sorted based on their word count similarities. We see clusters of highly-similar labels, such as the top left corner with labels relating to *"card"*, and the middle cluster with labels relating tor *"top_up"*.